Time-series,
Spring, 2026

Seasonality, SARIMA structure, seasonal differencing, SARIMAX

Faculty of DS & AI
Spring semester, 2026

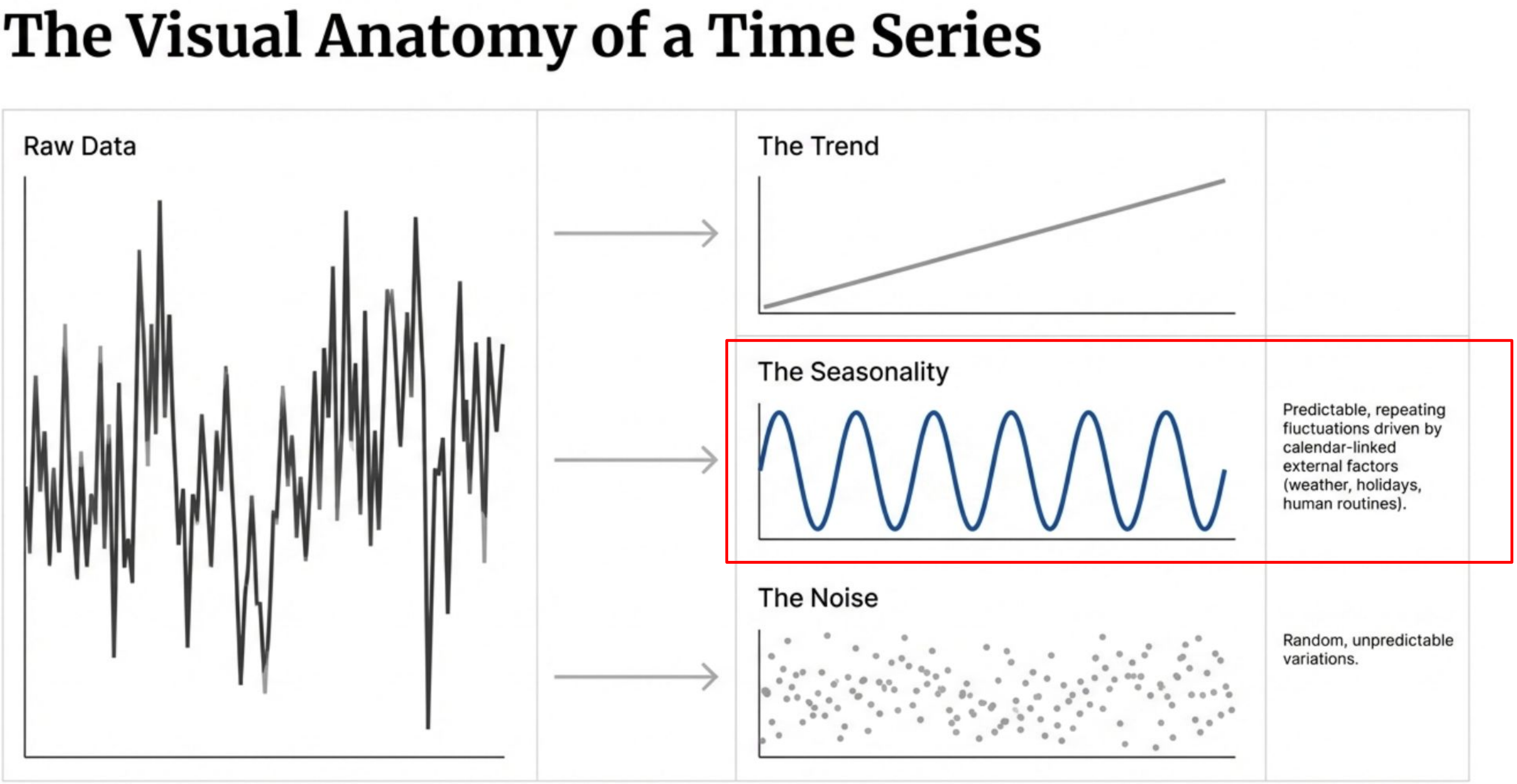
Trong-Nghia Nguyen

Business AI Lab

Content

- **Nature of Seasonality**
- Seasonal Differencing and Stationarity
- SARIMA
- SARIMAX

Nature of Seasonality



Nature of Seasonality

Defining the Seasonal Parameter (s)

The parameter s (or m) defines the exact number of observations within one complete seasonal cycle.

Monthly Data **s = 12**	Quarterly Data **s = 4**
Weekly Data (Daily Obs) **s = 7**	Hourly Data (Daily Cycle) **s = 24**

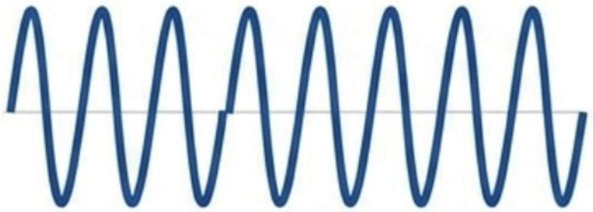
Nature of Seasonality

Typology of Seasonal Patterns

Annual Seasonality

Driven by weather or annual routines.

Example: Ice cream sales peaking uniformly during summer months.



Monthly / Weekly Seasonality

Driven by payroll cycles or weekend habits.

Example: Cinema ticket volumes surging predictably every Saturday and Sunday.

Holiday Seasonality

Extreme fluctuations around specific events (e.g., Lunar New Year, Black Friday).

Highly prone to creating model noise if holiday dates shift dynamically across calendar years.

Nature of Seasonality

Diagnostic Matrix: Seasonality vs. Cyclicity

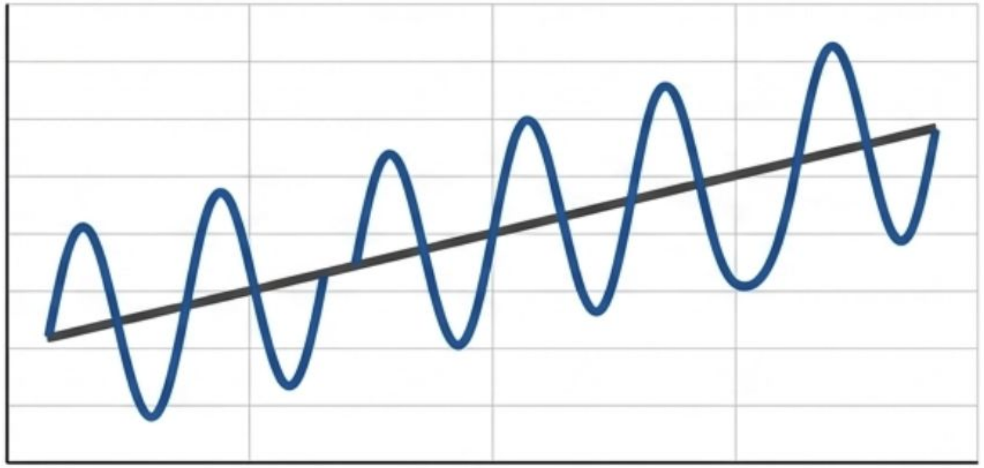
	Seasonality	Cyclicity
Frequency	Fixed and strictly known (e.g., exactly 12 months).	Variable and shifting (e.g., 2 to 10 years).
Primary Cause	Calendar constraints, weather, scheduled events.	Macroeconomic phases, boom/bust cycles.
Amplitude	Highly stable or predictably proportional to trend.	Highly volatile and difficult to forecast.
Classic Example	Mooncake sales surging exclusively in the 8th lunar month.	Real estate market valuations over a decade.

Nature of Seasonality

Equation to Chart Mapping: Structural Models

Additive Model

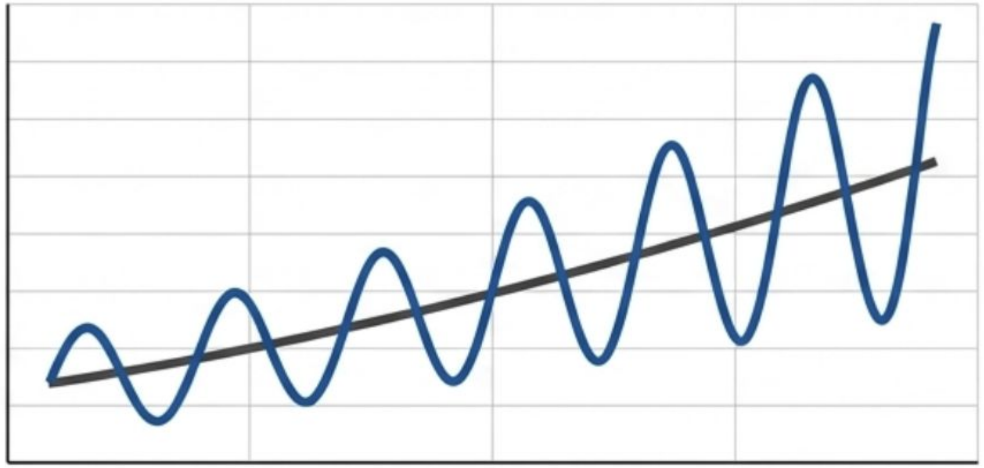
$Y_t = Trend_t + Seasonality_t + Residual_t$



Seasonal amplitude remains rigid regardless of trend direction. Used when **absolute error** variance is stable.

Multiplicative Model

$Y_t = Trend_t \times Seasonality_t \times Residual_t$



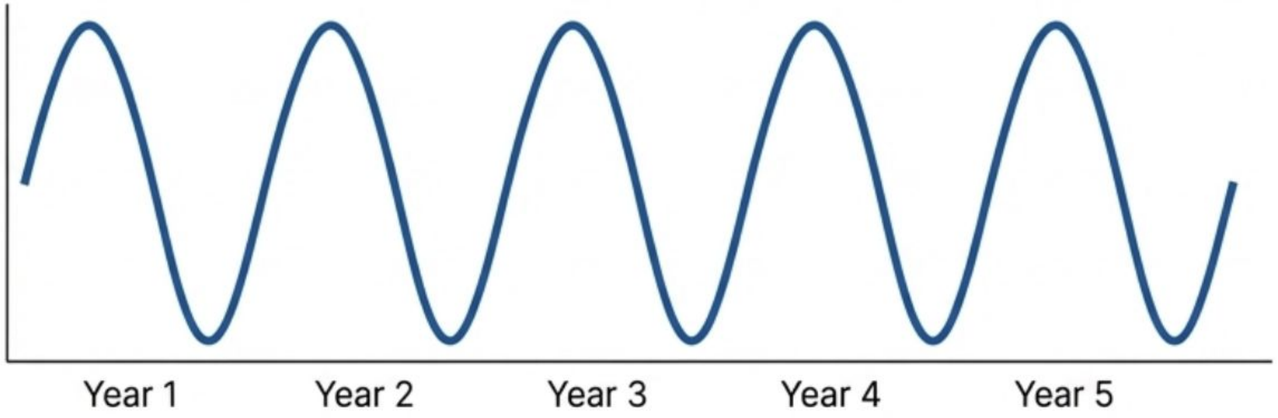
Seasonal amplitude **scales proportionally** with the trend. Dominant in retail and economics.

Nature of Seasonality

The Nature of the Signal: Deterministic vs. Stochastic

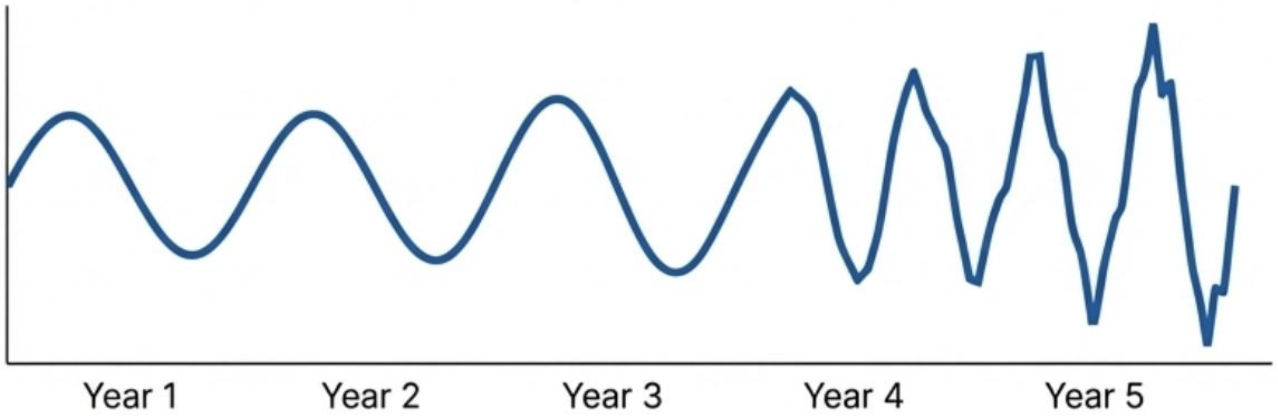
Deterministic Seasonality

The seasonal pattern repeats flawlessly without any deviation in timing or intensity across years. Resolvable using static Seasonal Dummy variables.



Stochastic Seasonality

The seasonal pattern undergoes evolutionary changes (e.g., shifting consumer shopping habits over a decade). Requires dynamic models like SARIMA to capture signal drift.



Nature of Seasonality

The Extraction Toolkit (Python statsmodels)

seasonal_decompose (Moving Average)

The legacy approach. Simple and highly interpretable.

Limitation: Assumes seasonality is perfectly static (Deterministic) and forces data loss at the extreme ends of the time series.

STL (Seasonal-Trend decomposition using LOESS)

The modern standard. Highly robust algorithm utilizing localized regression.

Advantages: Captures evolving (Stochastic) seasonality over time. Highly resistant to anomalous data spikes and outliers.

Nature of Seasonality

The Analytical Pipeline for Decomposition

1. Visual Inspection

Plot a simple Line chart to identify obvious recurring peaks/troughs.

2. ACF Testing (Autocorrelation)

Verify statistical significance. Look for spikes crossing the threshold at strict lags of s, 2s, 3s...

3. STL Extraction

Apply the STL
Apply the STL algorithm to isolate the seasonal component from the raw signal.

4. Stationarity Check

Confirm that the residual series (post-extraction) has achieved Stationarity.

Nature of Seasonality

Synthesis: The Seasonality Cheat Sheet

The Parameters (s)

Monthly = 12 | Quarterly = 4 | Weekly = 7 | Hourly = 24

Model Selection Framework

- Constant wave height → **Additive ($Y = T + S + R$)**
- Expanding wave height → **Multiplicative ($Y = T \times S \times R$)**

↓

Apply Logarithm

Diagnostic Reminders

- Seasonality: Fixed frequency, calendar-driven, predictable.
- Cyclicity: Variable frequency, macro-driven, erratic.
- Extraction: Default to STL over Moving Average for stochastic signals and outlier resistance.

Content

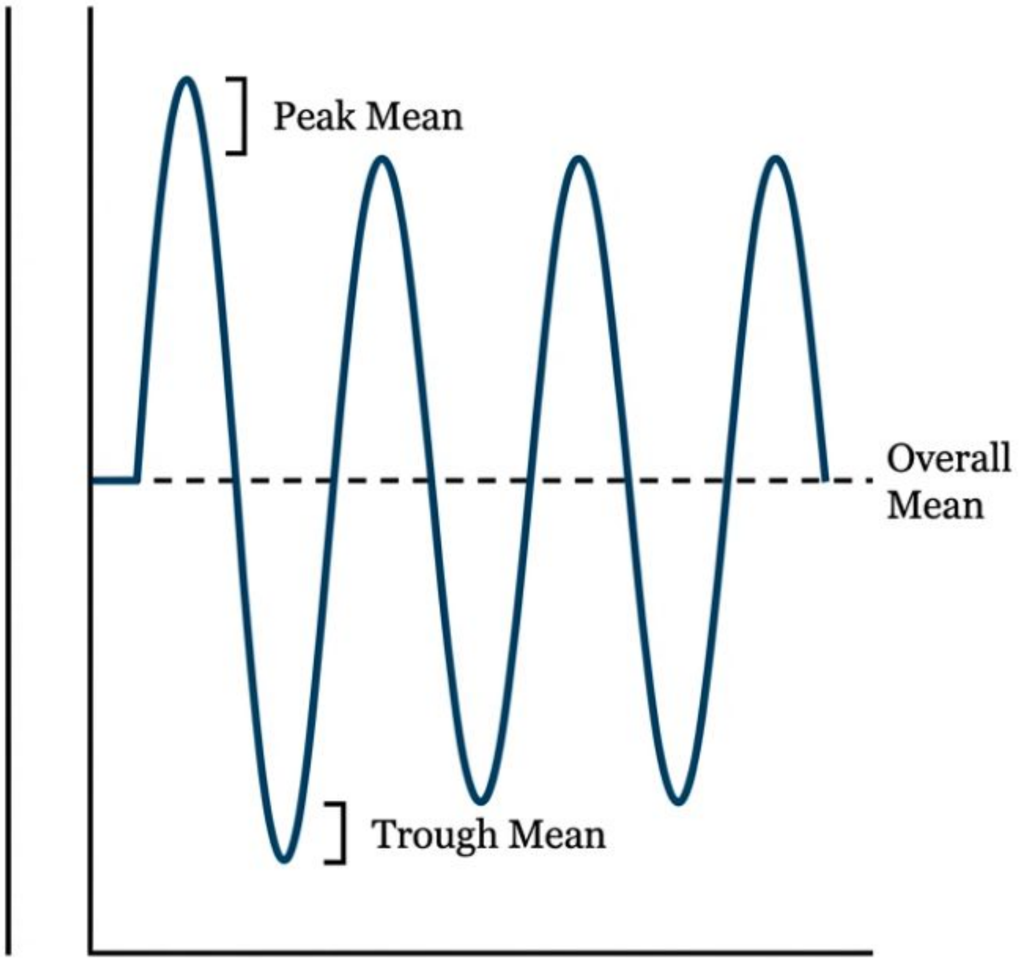
- Nature of Seasonality
- **Seasonal Differencing and Stationarity**
- SARIMA
- SARIMAX

Seasonal Differencing and Stationarity

Why Basic Stationarity is Insufficient

Core Concept: First-order differencing (d=1) successfully removes linear trends, but fails against a more subtle violation: Seasonal Non-stationarity.

The Phenomenon: Even if a sequence has no overall upward or downward trajectory (no trend), it remains non-stationary if the mean systematically shifts across repeating cyclic seasons.



Seasonal Differencing and Stationarity

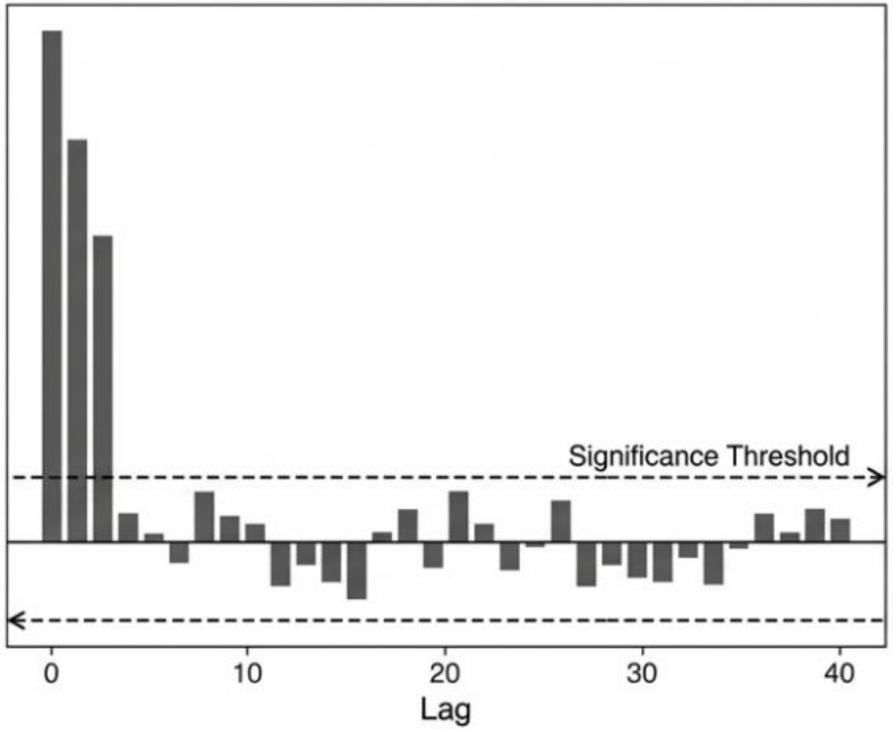
Diagnostic Matrix: Trend-Based vs. Seasonal-Based Non-Stationarity

Characteristic	Ordinary Time Series (Trend)	Seasonal Time Series
Cause	Mean changes continuously over time	Mean changes repeatedly in a cyclic pattern
ACF Signature	Gradual decay starting from Lag 1	Distinct spikes repeating at multiples of season s (e.g., 12, 24, 36)
Objective	Flatten the slope of the trend line	Neutralize recurring peaks and troughs
Mechanism	$\mathbf{Y_t - Y_{t-1}}$ (non-seasonal differencing)	$\mathbf{Y_t - Y_{t-s}}$ (seasonal differencing)

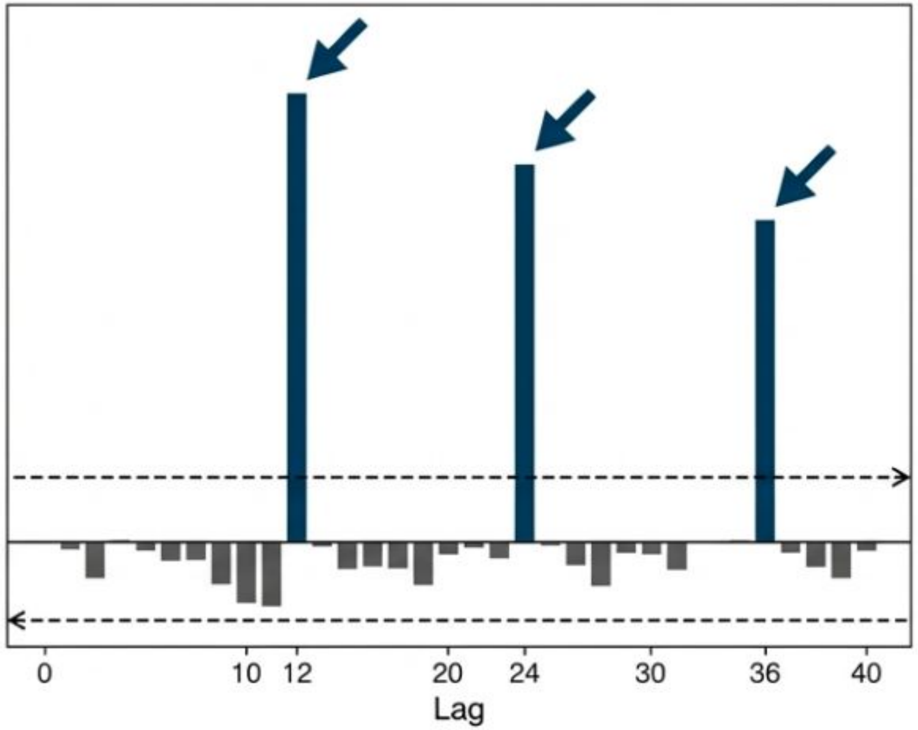
Seasonal Differencing and Stationarity

ACF Diagnostics: Distinguishing Signal Signatures

Non-Seasonal Spikes: Occur at early lags (1, 2, 3). Used to determine standard p, q parameters.



Seasonal Spikes: Spikes repeating strictly at multiples of the the season s. Slow decay of these multiples provides definitive proof that seasonal differencing (D=1) is required.



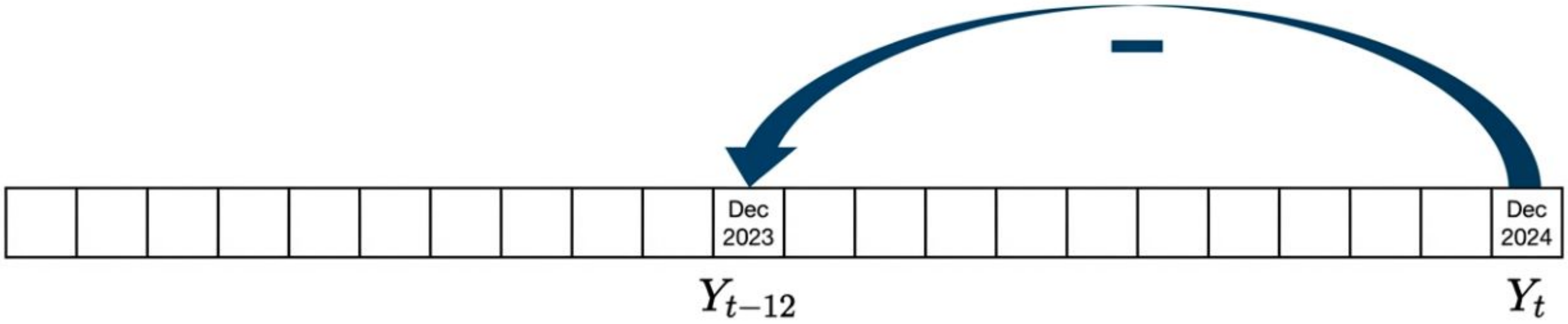
Seasonal Differencing and Stationarity

Seasonal Differencing (D=1)

Definition: Transformation achieved by calculating the difference between the current observation and the observation at the exact same point in the previous cycle.

Meaning: Eliminates the "average level" of the season. Example: If December sales are always artificially high due to holidays, subtracting last December's value neutralizes the seasonal effect.

$$\Delta_s Y_t = Y_t - Y_{t-s}$$

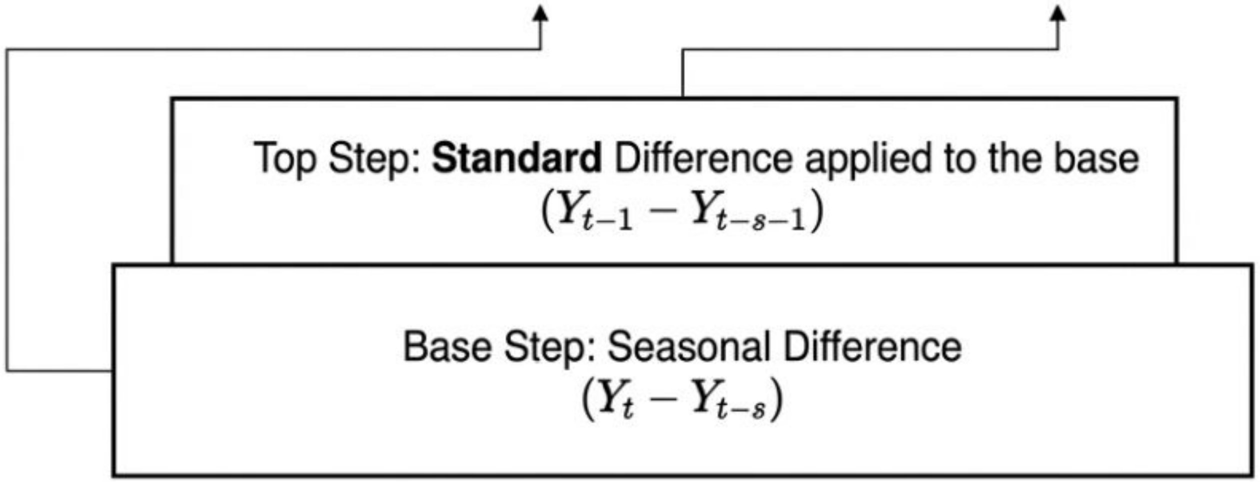


Seasonal Differencing and Stationarity

Mixed Differencing: Handling Trend + Seasonality

- The Reality: Datasets frequently exhibit both linear trends and seasonal fluctuations. A single seasonal difference may leave a residual trend or random walk structure behind.
- The Solution: Take the first-order difference of the seasonal difference. This measures the change in growth rate compared to the same period last year.

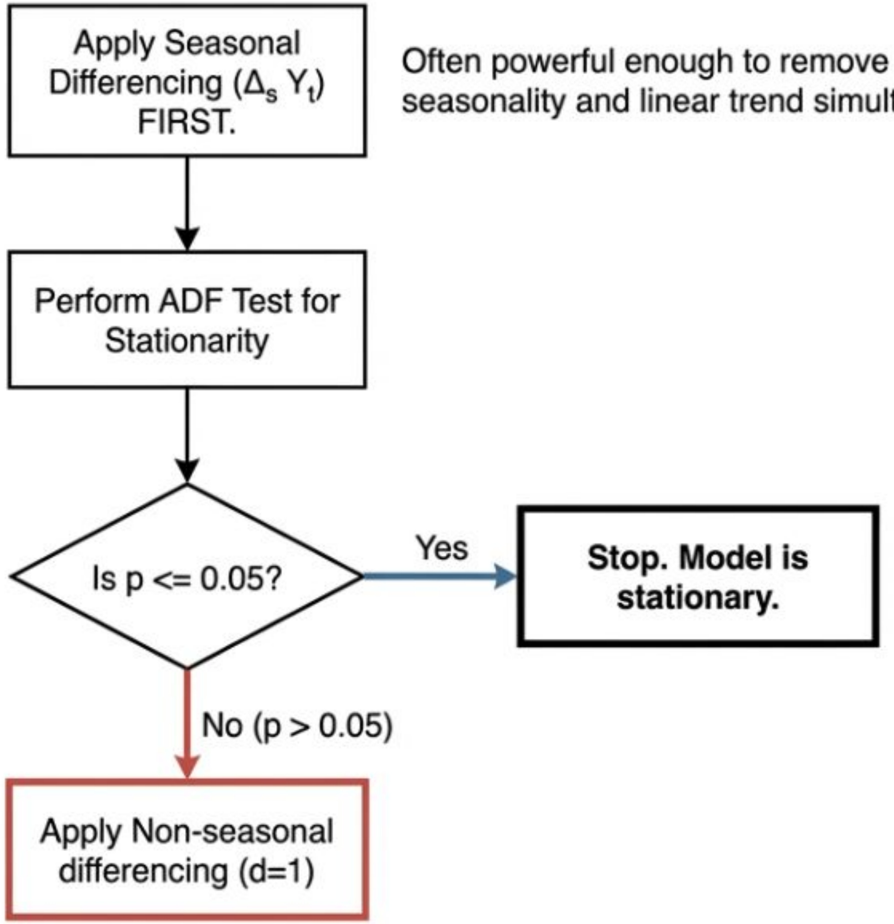
$$\Delta\Delta_s Y_t = (Y_t - Y_{t-s}) - (Y_{t-1} - Y_{t-s-1})$$



Seasonal Differencing and Stationarity

The Hyndman Protocol: Prioritizing Stability

Context: Professor Rob Hyndman's strategy to stabilize seasonal data without destroying underlying signal structures.

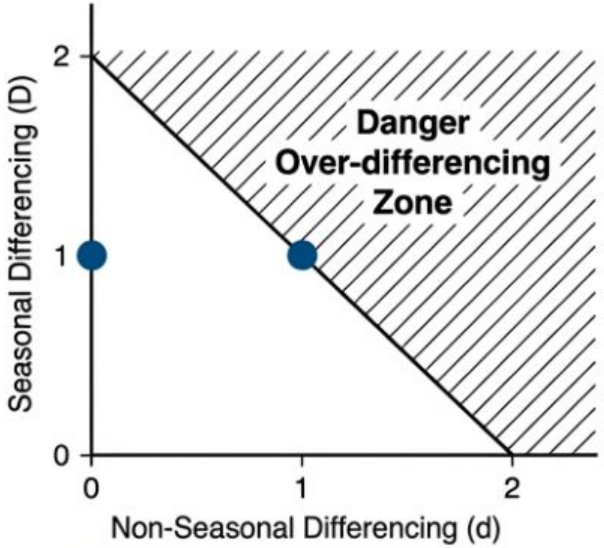


Seasonal Differencing and Stationarity

Framework Limits: The $d + D <= 2$ Rule

Core Rule: To prevent over-differencing, the total sum of non-seasonal (d) and seasonal (D) differencing orders should never exceed 2.

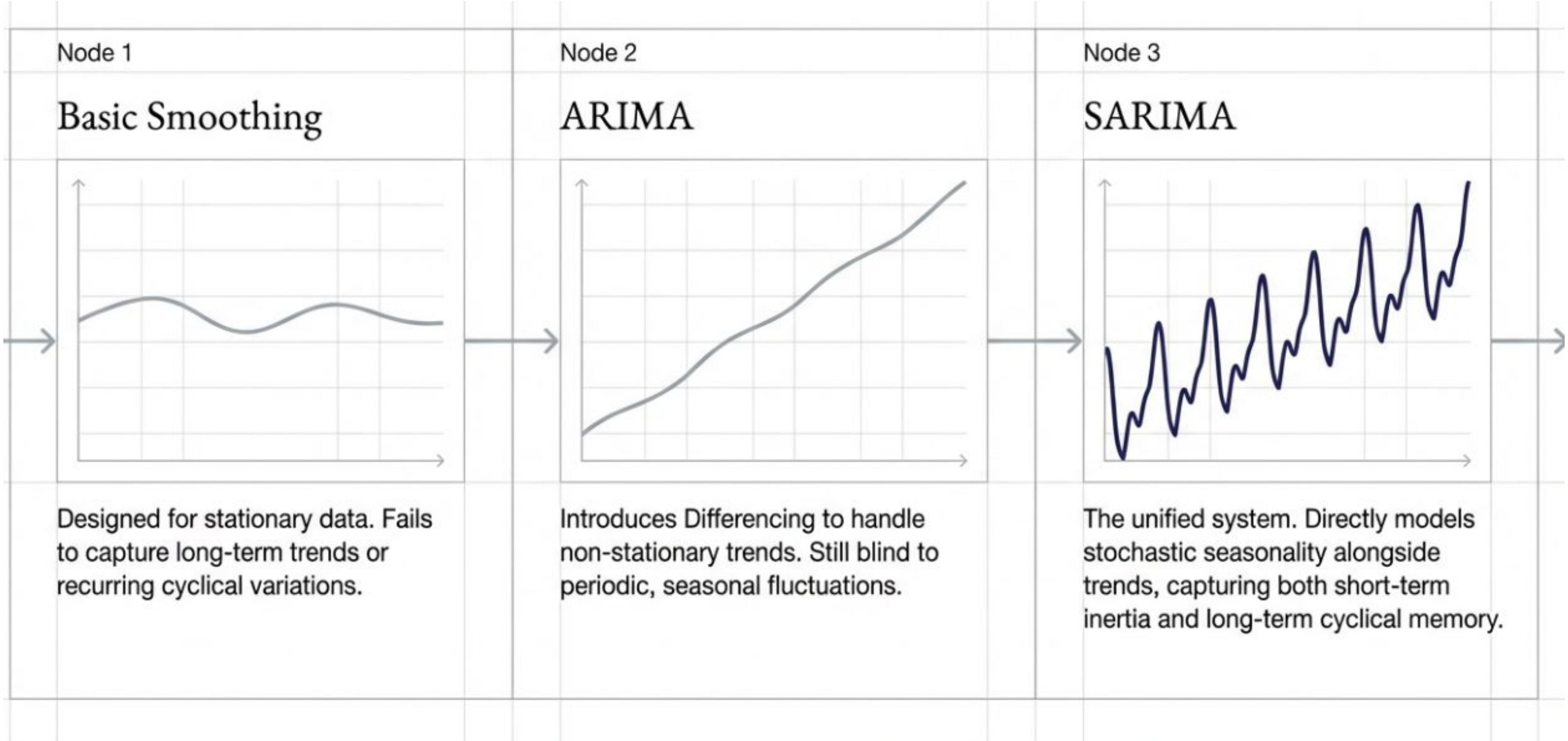
- (d=0, D=1): Strong seasonality, flat overall trend.
- (d=1, D=1): Strong seasonality, clear linear trend.
- (Rarely) (d=2, D=0): Exponential trend, no seasonality.



Content

- Nature of Seasonality
- Seasonal Differencing and Stationarity
- **SARIMA**
- SARIMAX

SARIMA


Node 1
Basic Smoothing
Designed for stationary data. Fails to capture long-term trends or recurring cyclical variations.
Node 2
ARIMA
Introduces Differencing to handle non-stationary trends. Still blind to periodic, seasonal fluctuations.
Node 3
SARIMA
The unified system. Directly models stochastic seasonality alongside trends, capturing both short-term inertia and long-term cyclical memory.

SARIMA

The Nature of Seasonality

Deterministic Seasonality

Static, unchanging patterns (e.g., selling exactly 500 extra units every Christmas).

Best Approach: Use dummy variables in regression.

Stochastic Seasonality

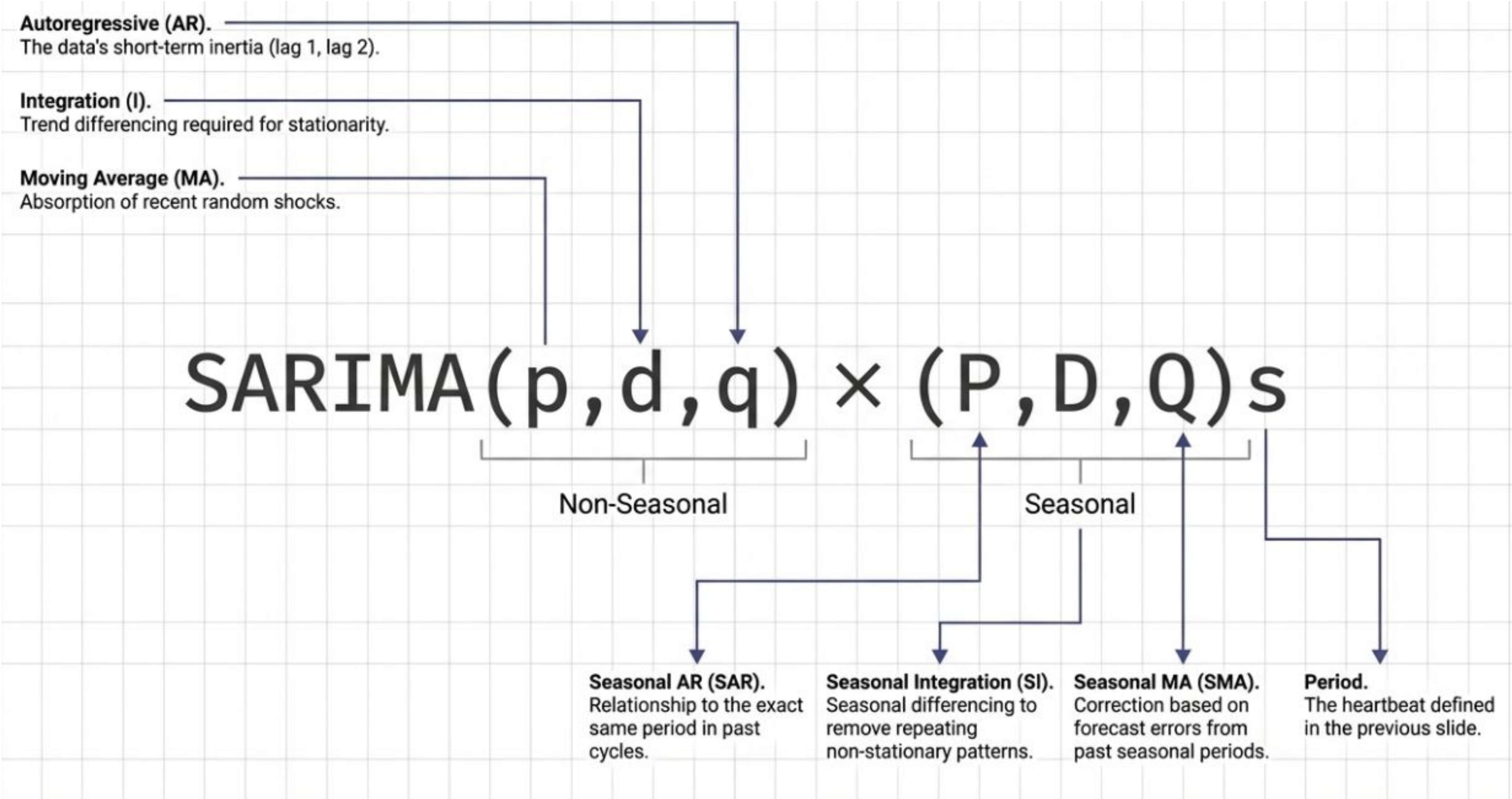
Evolving patterns where amplitude or timing shifts due to accumulated random shocks.

Best Approach: SARIMA (allows the seasonal component to evolve).

Defining the Pulse (s)

Data Frequency	Seasonal Period (s)	Practical Meaning
Hourly	$s = 24$	Daily cycle (hour of the day)
Daily	$s = 7$	Weekly cycle (day of the week)
Weekly	$s = 52$	Annual cycle (week of the year)
Monthly	$s = 12$	Annual cycle (month of the year)
Quarterly	$s = 4$	Annual cycle (quarter of the year)

SARIMA



Autoregressive (AR).
The data's short-term inertia (lag 1, lag 2).
Integration (I).
Trend differencing required for stationarity.
Moving Average (MA).
Absorption of recent random shocks.
SARIMA(p,d,q) × (P,D,Q)s
Non-Seasonal
Seasonal
Seasonal AR (SAR).
Relationship to the exact same period in past cycles.
Seasonal Integration (SI).
Seasonal differencing to remove repeating non-stationary patterns.
Seasonal MA (SMA).
Correction based on forecast errors from past seasonal periods.
Period.
The heartbeat defined in the previous slide.

SARIMA

Step 1: The Setup

Consider a monthly model (s=12) with one AR term and one SAR term.

$$(1 - \phi_1 B)(1 - \Phi_1 B^{12})y_t = \varepsilon_t$$

Step 2: The Expansion

Multiplying the polynomials reveals hidden layers of memory.

$$(1 - \phi_1 B - \Phi_1 B^{12} + \phi_1 \Phi_1 B^{13})y_t = \varepsilon_t$$

Step 3: The Forecasting Equation

Solving for y_t isolates the specific historical impacts.

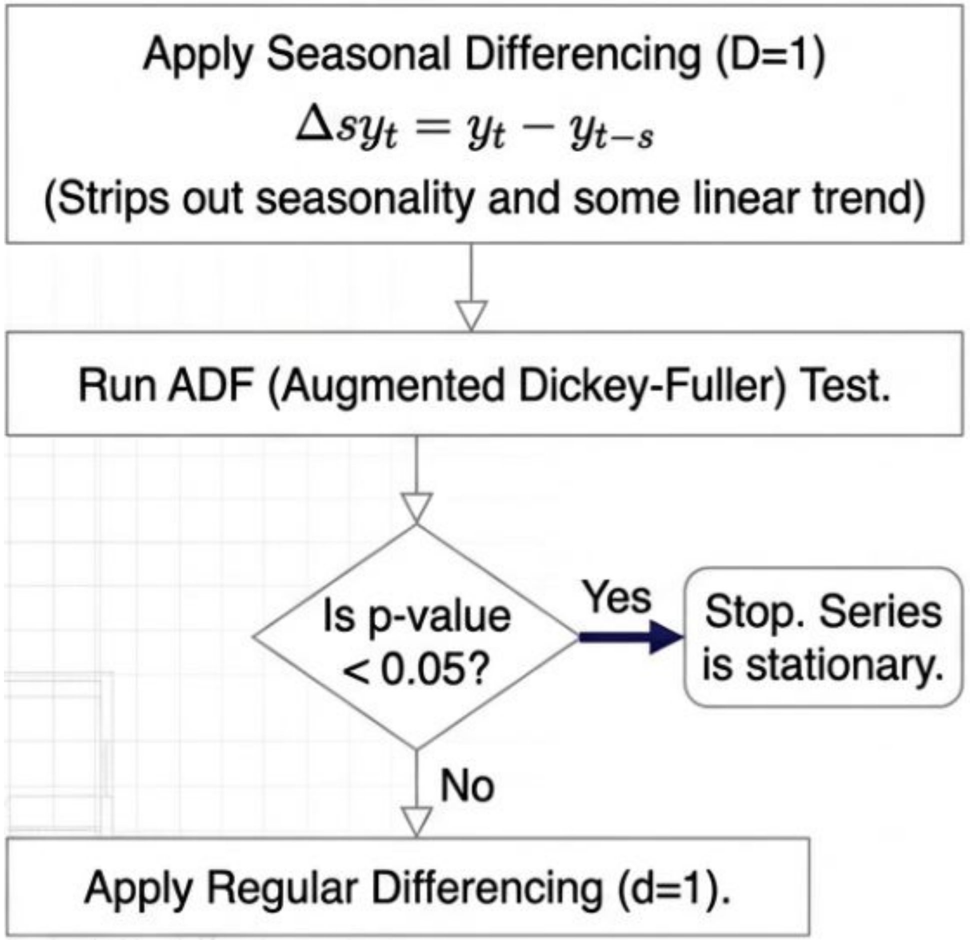
$$y_t = \phi_1 y_{t-1} + \Phi_1 y_{t-12} - (\phi_1 \Phi_1) y_{t-13} + \varepsilon_t$$

The Lag 13 Phenomenon

- $\phi_1\ y_{t-1}$ = Impact of the immediately preceding month.
- $\Phi_1\ y_{t-12}$ = Impact of the same month last year.
- $(\phi_1\ \Phi_1)\ y_{t-13}$ = **The Interaction Lag.** The impact of the month preceding the same month last year. This allows SARIMA to capture subtle shifts in seasonal patterns over time.

SARIMA

Hyndman's Protocol Flowchart

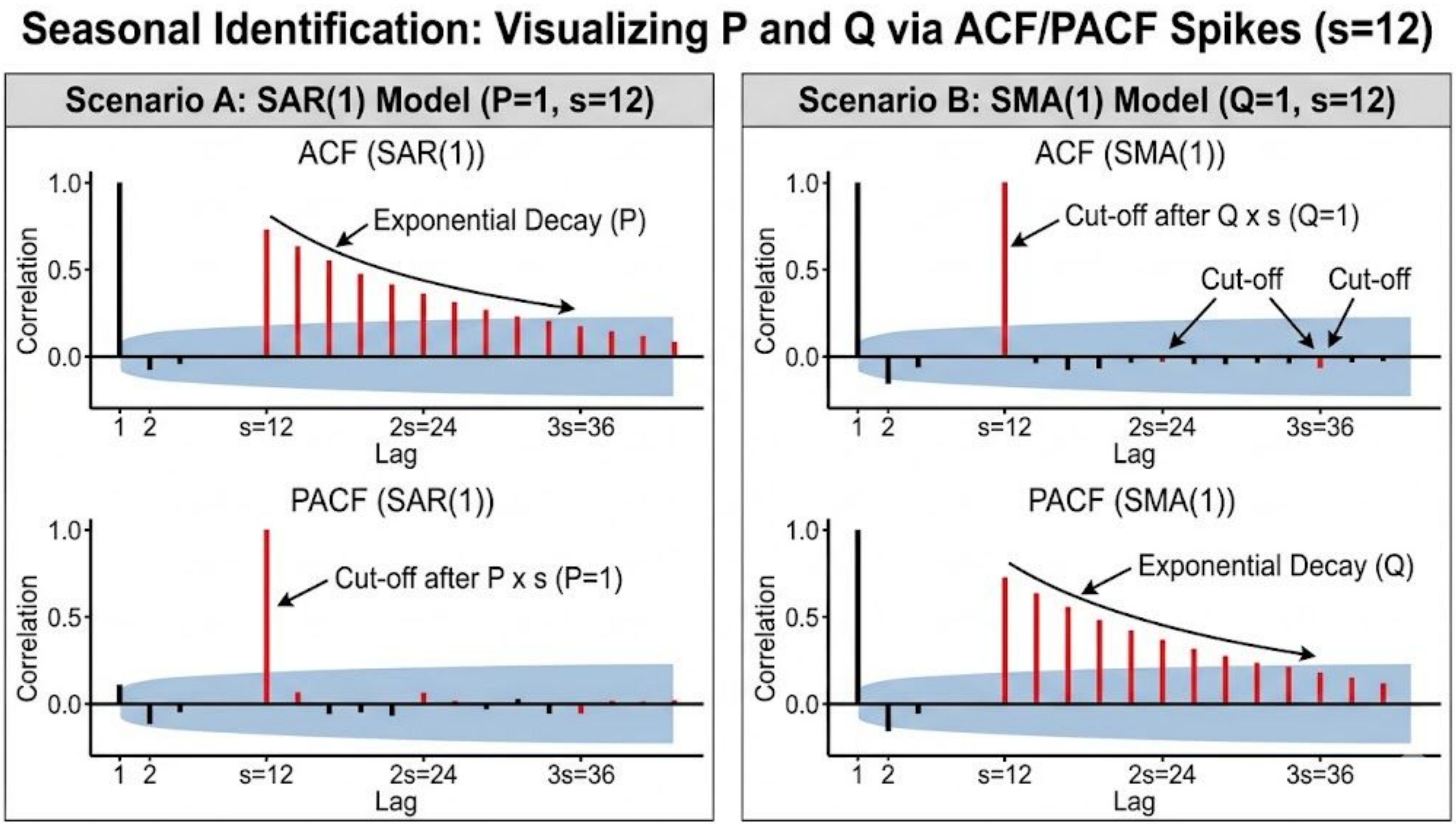


The Absolute Constraint

Rule: $d + D \leq 2$. Exceeding this causes over-differencing, inflating variance and creating artificial negative autocorrelation.

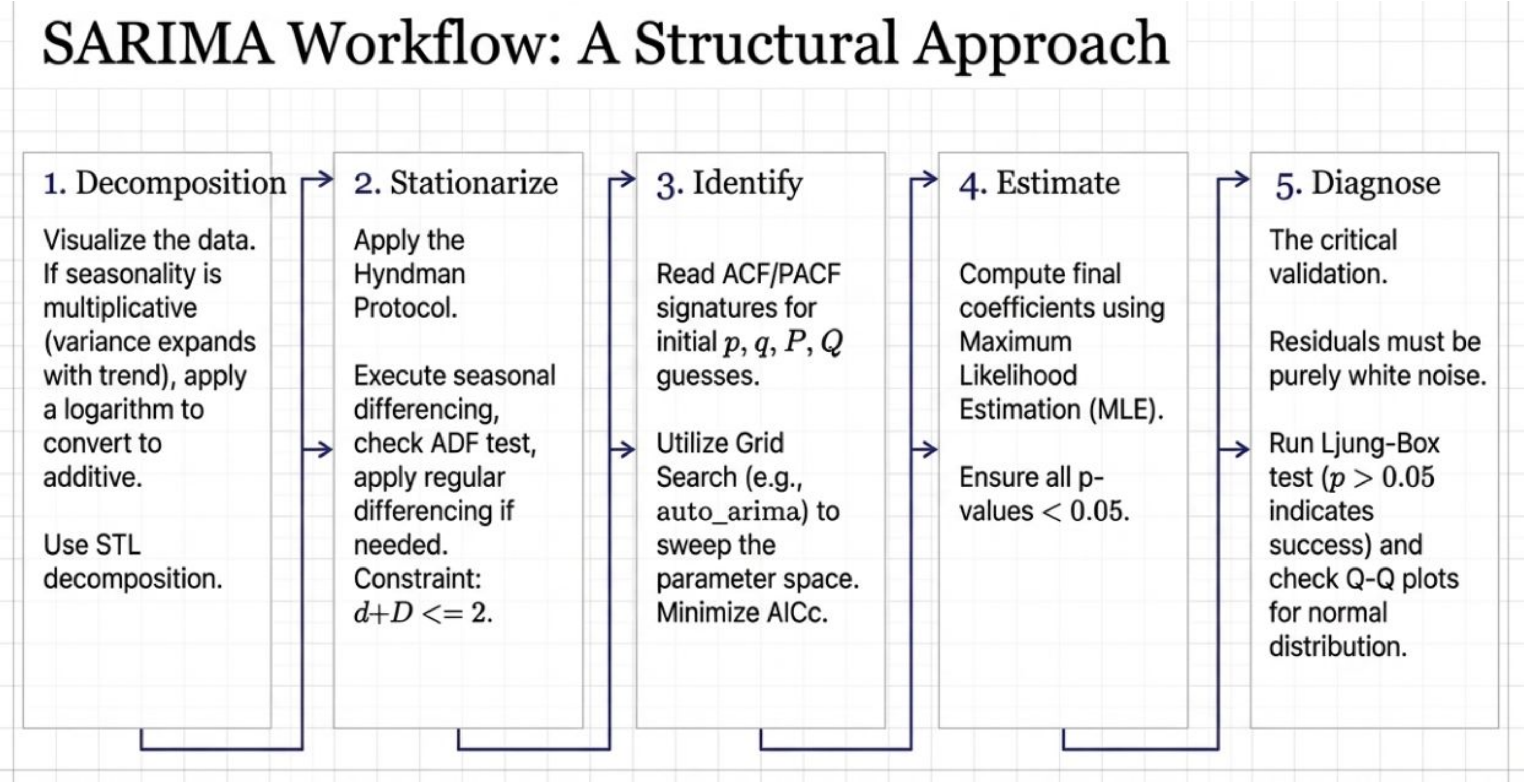
d \ D	0	1
0	Fully stationary data.	Linear trend, no seasonality.
1	Strong seasonality, stable mean.	Both trend and strong seasonality.

SARIMA



Seasonal Identification: Visualizing P and Q via ACF/PACF Spikes (s=12)
Scenario A: SAR(1) Model (P=1, s=12)
ACF (SAR(1))
Exponential Decay (P)
Correlation
1.0
0.5
0.0
1 2 s=12 2s=24 3s=36
Lag
PACF (SAR(1))
Cut-off after P x s (P=1)
Scenario B: SMA(1) Model (Q=1, s=12)
ACF (SMA(1))
Cut-off after Q x s (Q=1)
Cut-off
Cut-off
PACF (SMA(1))
Exponential Decay (Q)

SARIMA



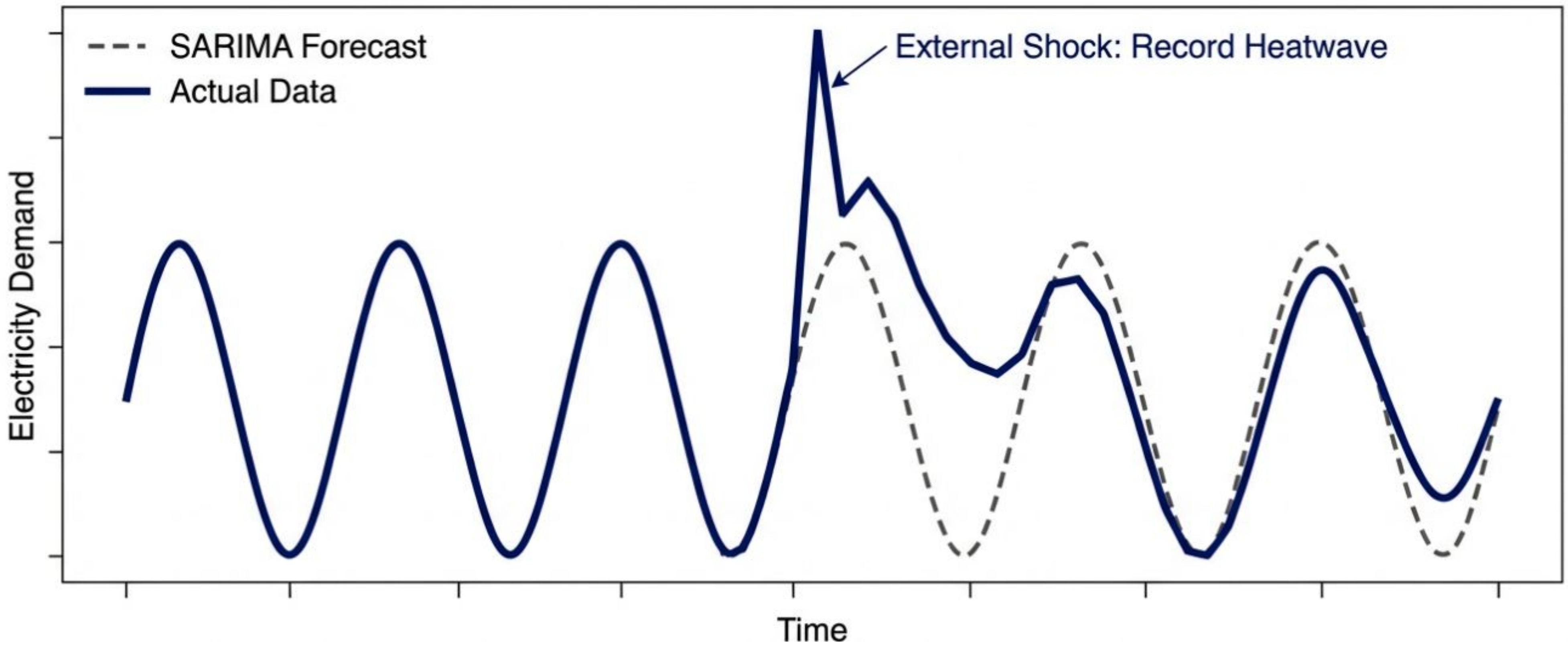
SARIMA Workflow: A Structural Approach
1. Decomposition
Visualize the data. If seasonality is multiplicative (variance expands with trend), apply a logarithm to convert to additive.
Use STL decomposition.
2. Stationarize
Apply the Hyndman Protocol.
Execute seasonal differencing, check ADF test, apply regular differencing if needed. Constraint: $d+D <= 2$.
3. Identify
Read ACF/PACF signatures for initial p, q, P, Q guesses.
Utilize Grid Search (e.g., auto_arima) to sweep the parameter space. Minimize AICc.
4. Estimate
Compute final coefficients using Maximum Likelihood Estimation (MLE).
Ensure all p-values < 0.05.
5. Diagnose
The critical validation.
Residuals must be purely white noise.
Run Ljung-Box test ($p > 0.05$ indicates success) and check Q-Q plots for normal distribution.

Content

SARIMAX

The Blind Spot of Internal Memory

Traditional SARIMA models are powerful at extracting internal rhythms (trends and seasonality). However, they exist in a vacuum. When real-world time series face sudden external shocks, models relying solely on past temporal data fail to react.



SARIMAX

Defining SARIMAX

Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors.

Key Takeaway Box:
SARIMAX solves the closed-system limitation by directly embedding external variables into the time-series memory, combining the structural precision of SARIMA with the explanatory power of linear regression.

SARIMAX

The Taxonomy of Exogenous Variables (X)

Independent variables that influence the target (Y) but are not influenced by Y within the system.

Quantitative Variables (Continuous)	Qualitative Variables (Dummies/Binary)
Temperature levels	Holidays
Oil prices	Lockdown periods
Inflation indices	Promotional events
Advertising expenditure	Black Friday events

SARIMAX

The Architectural Anatomy of SARIMAX

$$Y_t = \beta X_t + \eta_t$$

The Linear Regression Engine

The Regression Component. Captures the direct, immediate impact of external variables.

The Time Series Error Engine

The Residual Component. **Instead of assuming white noise, SARIMAX models these residuals using a full SARIMA** $(p, d, q)(P, D, Q)_S$ **process.**

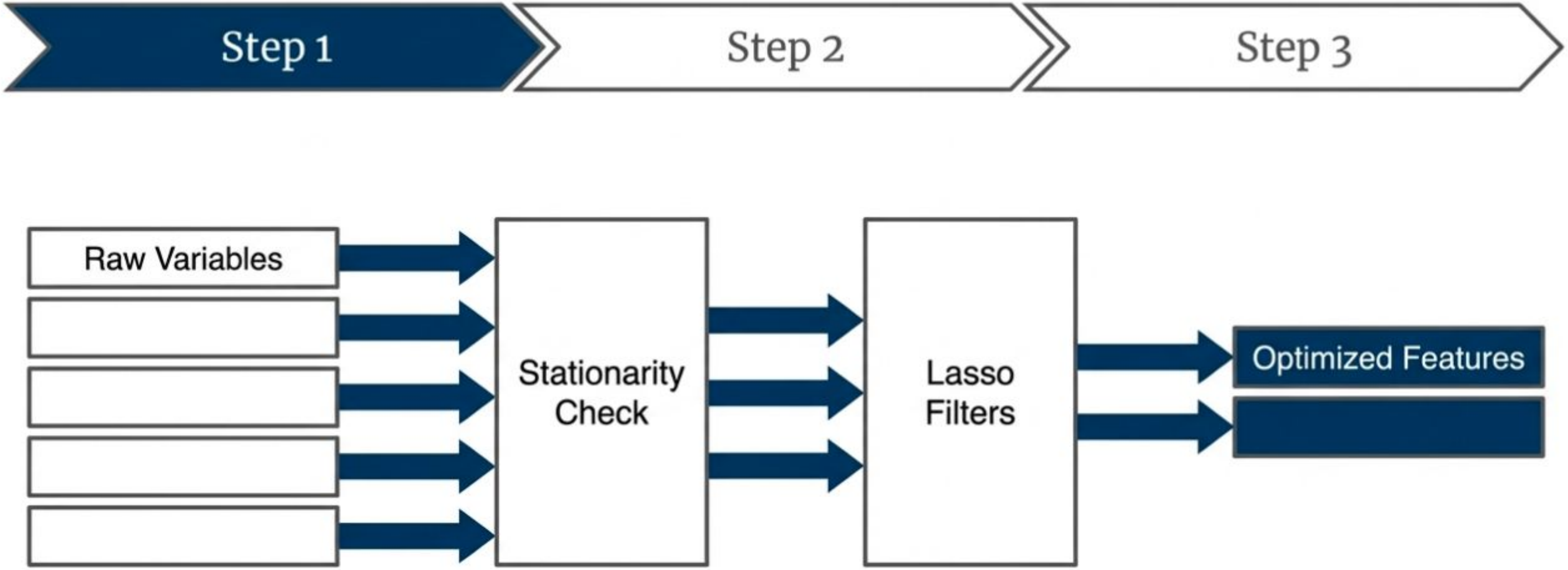
SARIMAX

Diagnostic Matrix: SARIMA vs. SARIMAX

	SARIMA	SARIMAX
Information Source	Relies exclusively on historical target data (Y).	Utilizes both historical target data and external predictors (X).
Response to Shocks	Suffers from “lagged realization” (must wait for temporal delay).	Achieves immediate reaction via external triggers.
Mathematical Assumption	Assumes the environment is a closed loop.	Assumes the environment is an open system driven by measurable external forces.

SARIMAX

Implementation Pipeline: 1. Selection & Processing



Stationarity Check: Exogenous variables (X) must be tested for stationarity. If non-stationary, apply differencing directly to X before integration.

Feature Selection: Guard against overfitting. Avoid blindly including all available variables. Utilize regularization techniques like Lasso Regression to filter for features with genuine explanatory power.

SARIMAX

Implementation Pipeline: 2. Estimation Mechanics

Step 1 | **Step 2** | Step 3

```
model = SARIMAX(endog=Y,
    exog=X_matrix,
    order=(p, d, q),
    seasonal_order=(P, D, Q, s))
```

Software Integration: In Python's statsmodels library, exogenous variables are passed as a matrix to the exog parameter alongside the target time series.

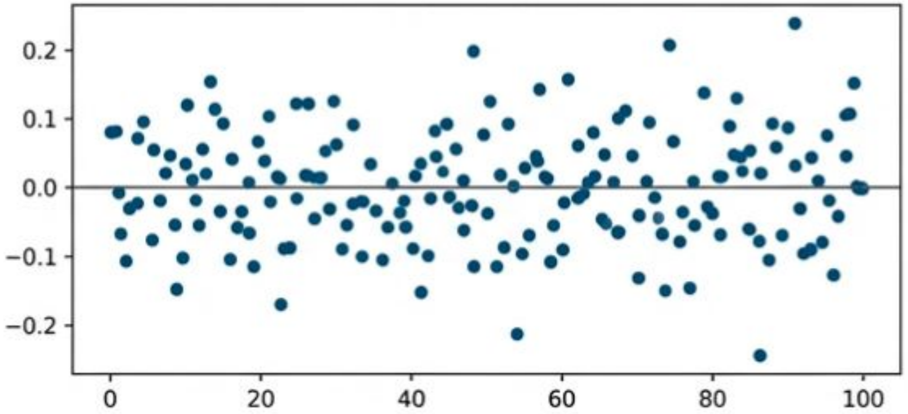
Optimization Target: Use the Corrected Akaike Information Criterion (AICc) to systematically evaluate and select the optimal $(p, d, q)(P, D, Q)_s$ parameter combination for the residual error structure.

SARIMAX

Implementation Pipeline: 3. Diagnostic Evaluation

Step 1 → Step 2 → **Step 3**

Diagnostic Checklist



Residual White Noise Check = PASS ✓

Residual Verification: Plot the residuals to guarantee they resemble pure white noise, indicating all temporal correlations have been successfully captured.

Diagnostic Checklist

	β Coefficient	p-value
1	Temperature	0.01
2	Advertising	0.08
3	Holiday	0.02

Statistical Significance: Rigorously audit the p-value of every exogenous coefficient (β). If a variable registers a p-value > 0.05, it lacks statistical significance and should be pruned to simplify the model.

Thank you